Classification of COVID-19 by SVM and CNN Using Chest X-Rays

Author: Akshay Mishra

Abstract

The severity of the impact of COVID-19 has been immense, it is putting great pressure on hospitals all over the world. People suffering from coronavirus suffer from inflammation in the lungs which damages the cells and tissues in the lungs. These damages can be discovered through analysis of frontal chest X-rays. In this paper, we compare the Support Vector Machine (SVM) and Convolutional Neural Network (CNN), ResNet-50, for the classification of COVID-19 vs Normal case using chest X-ray images. Both of the methods are tested on the same dataset. ResNet-50 achieves an accuracy of $96 \pm 2\%$ whereas naive SVM 90%. These methods are non-invasive and can be a predictor of COVID-19 in asymptomatic patients. Github: CS760 Final Project

1 Introduction

Appearing in 2002, Severe acute respiratory syndrome (SARS) was named the first pandemic of the 21st century. 8,098 people suffered from SARS in 26 countries with 774 fatalities [14]. COVID-19 has been proven far deadlier and virulent than SARS. As of Dec 12, 2020, the virus has spread to almost all nations in the world. As stated by WHO, there are now more than 69.1 million confirmed cases with over 1.5 million deaths all across the globe [18]. Table 1 shows the confirmed cases by geographical locations.

Country	*Total Cases*	*Total Death*	*Total Recovered*
USA	16,295,714	302,762	9,507,476
India	9,827,026	142,662	9,324,328
Brazil	6,836,313	180,453	5,954,745
Russia	2,597,711	45,893	2,059,840
Brazil	2,351,372	57,567	175,891

Table 1: Top 5 countries with most COVID-19 cases

In these tough times, researchers and scientists have been able to study and know a lot more about this virus and people infected by it. The symptoms of this virus begin like a typical viral, as it progresses it can cause nausea, shortness of breath, and can damage lungs [8]. In the early stages of COVID-19, ground glass patterns can be observed on the edges of the pulmonary vessels. These patterns can be visually identified through frontal chest X-ray images [10] [9]. Such identifications can only be made by radiologists. Additionally, around 17.9% the patients are asymptomatic, as seen in the case of Diamond Cruise ship [11]. Since a large proportion of the population suffering from the coronavirus might be asymptomatic, an early diagnosis is of real importance to provide the opportunity of immediate isolation. In this time when there are limited numbers of health practitioners versus the overwhelming number of COVID-19 cases, we need a quick and automatic method that can assist in classifying COVID-19 cases.

2 Related/Similar work

Machine learning techniques have long been used in the classification of images with binary as well as multi-class. SVMs were once popularly used for binary classification [16], recently there have been lots of advancement in the field of neural networks. So most of the work in this field has been under the same. CNNs for example have been used rigorously for the diagnosis of various diseases. Hartenstein et al. used deep learning to determine prostate cancer positivity from CT imaging [7].In addition, pre-trained CNNs have been able to successfully identify breast cancer [4], Alzheimer's disease[13], leukemia [6], pneumonia [17] etc.

Getting motivated by these papers recently there have been multiple papers that use pre-trained CNN to successfully predict COVID-19 using X-ray and CT images. One of the first studies, written by Apostolopoulos and Mpesiana, tried to detect COVID-19 cases based on chest X-ray images [2]. In this study, Apostolopoulos and Mpesiana use pre-trained CNNs such as MobileNet V2, VGG19, Inception, Xception, and Inception ResNet V2. Chest x-rays were obtained from two different datasets and were used to train the above networks. Based on their experiments MobileNet V2 achieved 97.40% accuracy and VGG19 achieved 98.7% accuracy for COVID-19 vs Normal classification. Similarly to compare different state-of-the-art CNNs Shi et al. performed a detailed review for computer-aided techniques for the detection of COVID-19 from X-ray and CT scans [1]. Castiglioni et al. and Narin et al. both use ResNet-50 for classification of COVID-19 and Normal cases on balanced dataset of 250-250 and 50-50 respectively [3] [12]. We selected ResNet-50 as our choice of pre-trained CNN network as well to compare it with the SVM.

3 Datasets and Methods

Both the method was implemented using COVID-19 public dataset and the Kaggle Pneumonia dataset. The COVID-19 dataset created by Cohen et al. consists of 950 chest X-rays and CT scans. For our study, we only focused on the chest X-rays, which were around 504 [5]. For the normal chest X-rays, we took data from Chest X-Ray Images (Pneumonia) dataset from Kaggle. This is a huge dataset with around 5800 images, we randomly sampled 504 normal images from this dataset to create a balanced 504-504 COVID-19 versus Normal X-ray study. We then split the data in 80:20 for train:test datasets, as shown in 2

As seen in Figure 1 provided by Cohen et al. [5] the radiologist marked the parts of the lungs that show damage due to COVID-19. Through our experiment we want the CNN model to learn to identify these regions and differentiate between COVID-19 and Normal scans, and the Support Vector Machine to create a decision boundary based on these dissimilarities between two cases. Below we talk more about the setting for each of the methods.

1. Support Vector Machine (SVM): "Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges [15] [16]. It is however mostly used in classification problems. In our current setup, we point each data item as a point in

Dataset	*Train*	*Test*	*Total*	*Link*
COVID-19	403	101	504	covid-chestxray-dataset
Normal	403	101	504	chest-xray-pneumonia
Combined	806	202	1008	

Table 2: Dataset's link and distribution used for this study

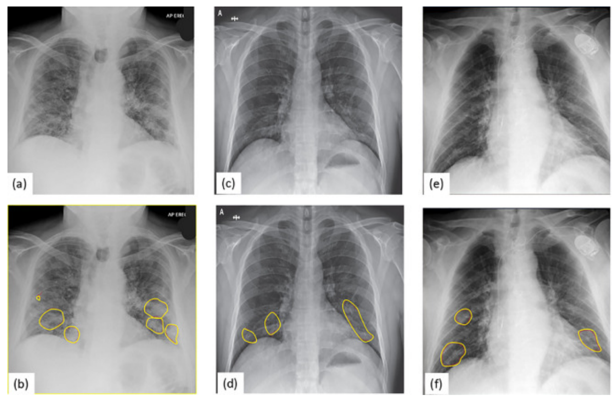

Figure 1: Three sample COVID-19 images, and the corresponding marked areas by radiologist

n-dimensional space with the value of each feature being the value of a particular coordinate, where n is the number number of features. As shown in Figure 2, we then perform classification by finding the hyper-plane that differentiates the two classes COVID-19 vs Normal. We used sklearn

The pre-processing step for SVM is straight forward. To create the features vectors (x) we first resized each image to 128X128 and then flatten it to numpy array. The label for each image acts as the output we want to predict (y), 0 for normal, and 1 for the COVID-19 cases. We then use scikit-learn package to fit the SVM classifier using the training dataset. The classifier has a few parameters like the kernel (we chose 'poly' to enable non-linear boundaries). The regularization parameter C, and gamma which define how far the influence of a single training point reaches, were tuned using a grid search to get the highest accuracy. Therefore, the best parameters for the SVM were {'C': 0.1, 'gamma': 1, 'kernel': 'poly'}. Once the model was fitted we did testing accordingly.

2. ResNet-50: It is difficult to obtain a huge number of labeled medical images. Especially in the case of COVID-19, the imaging dataset is steadily growing but is still far from enough for training a CNN from scratch. So, for more efficient training and easier gradient flow, we used a pre-trained ResNet-50 model. ResNet is one of the most popular CNN with the core idea of identity shortcut connection that skips one or more layers. This helps early layers in the networks to learn the dataset quickly, making the gradient updates for those layers much easier. The overall block diagram of the ResNet-50 model, and how it is used for COVID-19 detection is illustrated in Figure 3.

   As discussed before, we want the ResNet model to learn the differences shown in Figure 1. To do that, during the training phase we randomly rotate all the images by 10, since not all the images in the dataset are completely straight. Additionally, we resize all the images to 256X256. The training parameters of the model are as follows: learning rate of 1e-5, batch size of 16, and the number of epochs is 8. Additionally, a learning rate decay of 0.995 is applied to each step for optimization. We used

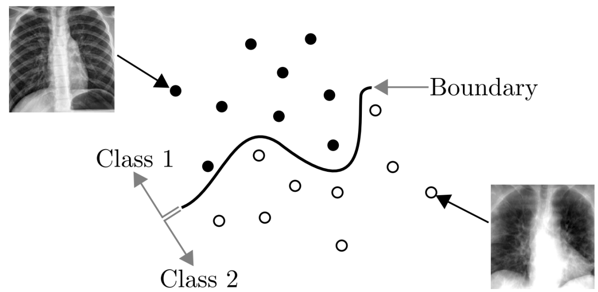

Figure 2: SVM classifies each point according to the region where it is located with respect to a boundary (Lecture notes SVM)

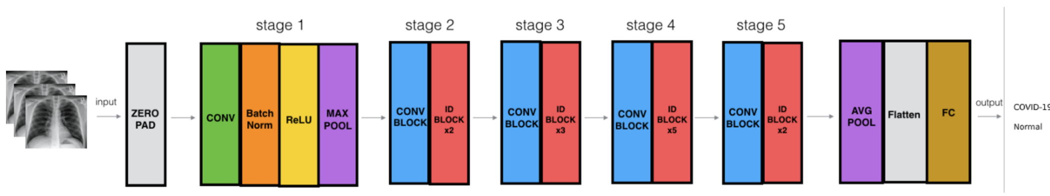

Figure 3: ResNet-50 model architecture

Pytorch as the machine learning framework to train the model. This model was trained on GEFORCE GTX 1080 with 10GB memory on a Linux system, which takes around 9 minutes to finish the training process.

4 Results

We tested both of the models on the same dataset. As seen in Table 2, we kept the number of images of COVID-19 and Normal cases the same. Additionally, we have an 80:20 split for training to testing cases. The evaluation criteria of both of the models are the same as well. More definitively, we wanted to test the precision, recall, and f1-score for both of the models as it gives us a sense of accuracy for both positive and negative cases.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{1}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{2}$$

$$Recall = 2. \frac{precision \, . \, recall}{precision + recall} \tag{3}$$

ResNet-50: This model predicts a probability score for each being a COVID-19 case or not. Higher the probability more confident the model is on the prediction. We set the regular threshold of 0.5 so that we can get a binary label for prediction. Figure 4 shows accuracy and loss scores per epoch for training and testing datasets. From the accuracy graph, we can see that the model quickly pick up the pattern in the images with an accuracy of 92% as it increases slowly up to 96%, whereas the loss approaches zero at the beginning of 8^{th} epoch. For this model, it performs really good for all three evaluation criteria, with a precision of 0.94, recall of 0.99, and f1-score of 0.96 (Table 3). To check the interpretability of the model we plotted guided-gradcam and gradcam outputs in Figure 5 and Figure 6. The CAM outputs are produced by up-sampling 8X8 sized features to 256X256. To make the significant regions pop we thresholded the colored region for easy visualization. The guided-gradcam provides more granularity over pixels affecting the model prediction. The gradcam images are up to interpretation.

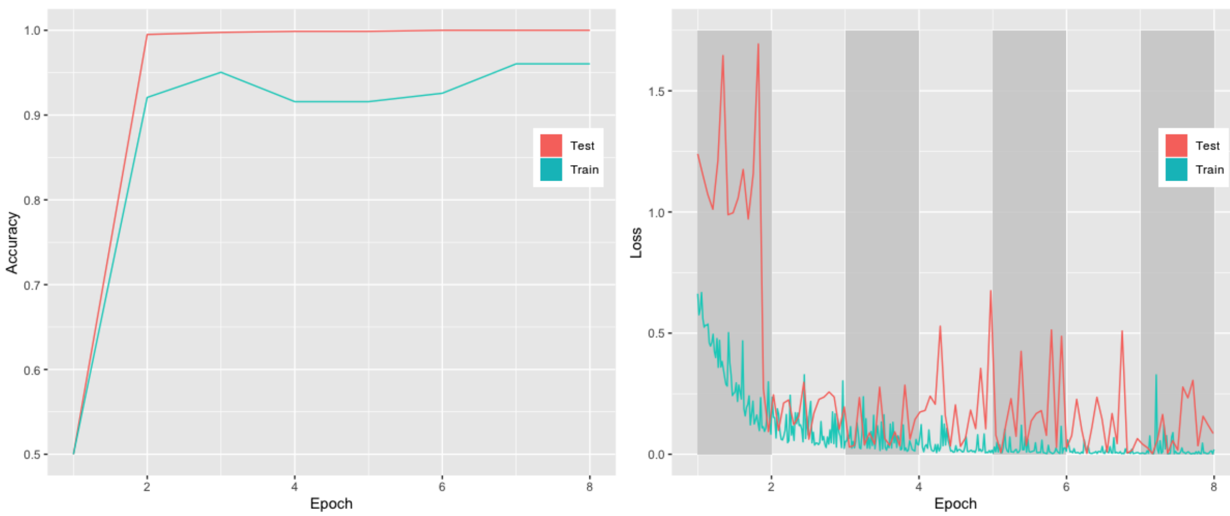

Figure 4: Left: The accuracy of the test and train datasets vs number of epoch. Right: The loss score graph on the test and train datasets, the alternate shades show the steps in each epoch

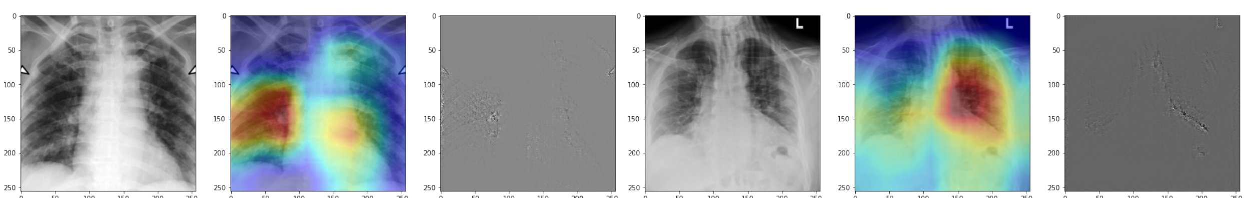

Figure 5: In the above heatmaps, we plot grad-cam and guided-gradcam for the cases in which model correctly predicts COVID-19 cases.

SVM: This model performed better than we expected. With an accuracy of 0.9 that is the total number of correct labels predicted divided by the size of the testing dataset. SVM was able to come up with a non-linear boundary to distinguish between the two classes in the dataset. From Table 3 we can see that it achieved precision value of 0.91, recall of 0.90, and f1-score of 0.96. Additionally, the model fitting time happens in around 10 seconds as well.

Both of these models produced an accuracy of greater than and equal to 90% same can be said about the precision, recall, and f1-scores. When comparing both these models among themselves it is clear that ResNet-50, which is a CNN, outperforms the results of SVM by a lot. Making ResNet-50 a superior method among both.

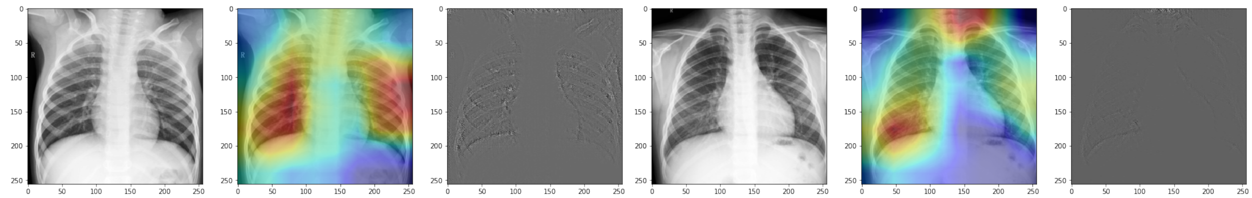

Figure 6: In the above heatmaps, we plot grad-cam and guided-gradcam for the cases in which model correctly predicts Normal cases.

Method	*Precision*	*Recall*	*F1-score*
SVM	0.91	0.90	0.90
ResNet-50	0.94	0.99	0.96

Table 3: Precison, recall, and f1-score for SVM and ResNet-50 for predicition of COVID-19

5 Conclusions and Future Work

X-ray images have been used in the diagnosis of medical conditions for a long time. When COVID-19 spread all over the world one of the earliest indicators of a patient being infected was through the diagnosis of chest X-ray images. Although, a limited number of radiologist experts and a skyrocketing amount of infected cases puts an immense amount of pressure on the hospitals. Additionally, even now the rapid antigen testing is not available in most of the small cities making the waiting time to get the test results anywhere between 24-72 hours. A reliable, quick, and automated algorithm for this task of distinction of COVID-19 cases from non-COVID-19 can speed up the early quarantining process. ResNet-50 model when compared to SVM crushed this classification task with $96 \pm 2\%$ accuracy. Having said that, one should keep in mind that this model should only be used as an early indicator so that precautionary measures that be taken as soon as possible. This should not be used as a diagnostic tool, as the motive of this project was to given an early start to the quarantining process in asymptomatic patients.

For future work, it would be interesting to see how a radiologist interprets the ResNet-50's grad cam output. Additionally, as the number of datasets increases revisiting the same problem with more data points should give better results. We would also want to test if the features extracted from ResNet-50 when put into Support Vector Classifier give better results than both of these models.

References

[1] Abraham, B., & Nair, M. S. (2020). *Computer-aided detection of COVID-19 from X-ray images using multi-CNN and Bayesnet classifier*. Biocybernetics and biomedical engineering, 40(4), 1436-1445.

[2] Apostolopoulos, I. D., & Mpesiana, T. A. (2020). *Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks*. Phys Eng Sci Med 43: 635–640.

[3] Castiglioni, I., Ippolito, D., Interlenghi, M., Monti, C. B., Salvatore, C., Schiaffino, S., ... & Sardanelli, F. (2020). *Artificial intelligence applied on chest X-ray can aid in the diagnosis of COVID-19 infection: a first experience from Lombardy, Italy*. medRxiv.

[4] Celik, Y., Talo, M., Yildirim, O., Karabatak, M., & Acharya, U. R. (2020). *Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images.* Pattern Recognition Letters.

[5] Cohen, J. P., Morrison, P., Dao, L., Roth, K., Duong, T. Q., & Ghassemi, M. (2020). *Covid-19 image data collection: Prospective predictions are the future.* arXiv preprint arXiv:2006.11988.

[6] Doan, M., Case, M., Masic, D., Hennig, H., McQuin, C., Caicedo, J., ... & Jamieson, D. (2020). *Label-free leukemia monitoring by computer vision.* Cytometry Part A, 97(4), 407-414.

[7] Hartenstein, A., Lübbe, F., Baur, A. D., Rudolph, M. M., Furth, C., Brenner, W., ... & Penzkofer, T. (2020). *Prostate Cancer Nodal Staging: Using Deep Learning to Predict 68 Ga-PSMA-Positivity from CT Imaging Alone.* Scientific Reports, 10(1), 1-11.

[8] Hui, D. S., Azhar, E. I., Madani, T. A., Ntoumi, F., Kock, R., Dar, O., ... & Zumla, A. (2020). *The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health-The latest 2019 novel coronavirus outbreak in Wuhan, China.*International Journal of Infectious Diseases, 91, 264-266.

[9] Kanne, J. P., Little, B. P., Chung, J. H., Elicker, B. M., & Ketai, L. H. (2020). *Essentials for radiologists on COVID-19: an update—radiology scientific expert panel.*

[10] Minaee, S., Kafieh, R., Sonka, M., Yazdani, S., & Soufi, G. J. (2020). Deep-covid: *Predicting covid-19 from chest x-ray images using deep transfer learning.* arXiv preprint arXiv:2004.09363.

[11] Mizumoto, K., Kagaya, K., Zarebski, A., & Chowell, G. (2020). *Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020.* Eurosurveillance, 25(10), 2000180.

[12] Narin, A., Kaya, C., & Pamuk, Z. (2020). *Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks.* arXiv preprint arXiv:2003.10849.

[13] Oh, K., Chung, Y. C., Kim, K. W., Kim, W. S., & Oh, I. S. (2019). *Classification and visualization of Alzheimer's disease using volumetric convolutional neural network and transfer learning.* Scientific Reports, 9(1), 1-16.

[14] SARS. (2005, May 03). Retrieved December 12, 2020, from https://www.cdc.gov/sars/about/faq.html

[15] Scholkopf, B. (1998). Support vector machines: a practical consequence of learning theory. IEEE Intelligent systems, 13.

[16] Stanevski, Nikolay & Tsvetkov, Dimiter & CLASSIFIER, MARGIN. (2005). *Using Support Vector Machine as a Binary Classifier.*

[17] Varshni, D., Thakral, K., Agarwal, L., Nijhawan, R., & Mittal, A. (2019, February). Pneumonia detection using CNN based feature extraction. In 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1-7). IEEE.

[18] WHO Coronavirus Disease (COVID-19) Dashboard. (n.d.). Retrieved December 12, 2020, from https://covid19.who.int/

[19] Yoon, H., Lee, J., Oh, J. E., Kim, H. R., Lee, S., Chang, H. J., & Sohn, D. K. (2019). *Tumor identification in colorectal histology images using a convolutional neural network.* Journal of digital imaging, 32(1), 131-140.